Chapter 12

A NEW PROCESS MODEL FOR TEXT STRING SEARCHING

Nicole Beebe and Glenn Dietrich

Abstract Investigations involving digital media (e.g., hard disks and USB thumb drives) rely heavily on text string searches. Traditional search approaches utilizing matching algorithms or database technology and tree-based indexing algorithms result in an overwhelming number of "hits" – a large percentage of which are irrelevant to investigative objectives. Furthermore, current approaches predominantly employ literal search techniques, which lead to poor recall with respect to investigative objectives. A better approach is needed that reduces information retrieval overhead and improves investigative recall. This paper proposes a new, high-level text string search process model that addresses some of the shortfalls in current text string search paradigms. We hope that this model will stimulate efforts on extending information retrieval and text mining research to digital forensic text string searching.

Keywords: Text string search, information retrieval, text mining, process model

1. Introduction

The digital forensics discipline is experiencing renewed attention at a time when data storage requirements and capabilities are increasing steeply. Hinshaw [10] reports that corporate data storage is doubling every nine months – twice as fast as Moore's Law. Because of this growth, it is not uncommon for larger corporations and law enforcement agencies to face digital investigations involving data sets of a terabyte or more in size [18, 23].

Current digital forensic tools are incapable of handling large data sets in an efficient manner [20]. Their overall efficiency is constrained by their reliance on relatively simple hashing and indexing algorithms. Most forensic tools and processes are not scalable to large data sets [4, 8, 21]. Even with moderately large (200 GB) data sets, data extraction

and analysis become inordinately slow and inefficient. Processing times for limited keyword searches (10-20 keywords) can take a day or more, and human analysts are overwhelmed by the number of hits to review.

Digital investigations are also hindered by the limited cognitive processing capability of human analysts. As data sets increase in size, the amount of data required for examination and analysis also increases. This obviates the investigator's ability to meticulously review all keyword search hits, all files by file type, or all applicable system logs. It is, therefore, imperative that the digital investigation process be improved.

Digital investigation processes and tools under-utilize computer processing power through continued reliance on simplistic data reduction and mining algorithms. The analytical burden was shifted to human analysts at a time when human labor was cheap and computers were expensive. For quite some time, however, conditions have been reversed, but the digital forensics field has continued to levy the preponderance of the analytical burden on human analysts.

Digital forensics is not the only discipline faced with the task of sifting through and drawing conclusions from massive volumes of data. Other disciplines have employed data mining and information retrieval techniques to solve this problem. However, little research has focused on applying these techniques to criminal forensics, and even less to digital forensics. This paper explores digital forensics from a systems perspective in order to identify the major constraints and propose solutions. Specifically, the text string search process applied during digital forensic investigations is explored, and a new, high-level, multi-algorithmic approach is proposed in the form of a high-level, theoretical process model.

2. Background

Digital investigations are categorized as network forensic investigations or media forensic investigations. Mukkamala and Sung [16] define network forensics as "analyzing network activity in order to discover the source of security policy violations or information assurance breaches." Media forensics, on the other hand, involves the analysis of digital media in order to confirm or refute allegations and/or obtain information. Such allegations may be civil or criminal in nature; and the subsequent investigation may be forensic or non-forensic – the difference being whether investigative goals involve judicial action.

Media forensics largely involves four primary investigatory digital search tactics: text string search, file signature search and analysis, hash analysis, and logical data review. Text string search (string matching, pattern matching or index-based searching) is designed to scan the digi-

tal media at the physical level (independent of logical data structures, file allocation status, partitioning, etc.) to locate and analyze data wherein a specific text string or pattern is located. File signature search/analysis is designed to search the digital media at the physical level to locate files by file type, without relying on potentially falsified file extensions at the logical level or the absence of them at the physical level (e.g., the file is no longer stored logically). Hash analysis is designed to find files identical to a known exemplar or to eliminate files that are "known goods" – that is to say, they are standard system and/or application files known to be irrelevant to the investigation. Finally, logical data review is designed to examine known repositories of logically stored data with potentially probative value (e.g., temporary Internet files, Internet history, registry and "My Recent Documents").

These tactics are employed for different reasons. While the tactics may produce redundant information, each tends to contribute unique information to the investigation. As a result, all four tactics are often employed during an investigation to obtain all relevant information.

This paper focuses on text string search. Text string searches are relied upon heavily by digital forensic investigators, as readable text and text-based documents are important artifacts in most investigations (e.g., email, web browsing history, word processing documents and spreadsheets). However, digital forensic text string searching is prone to high false positive and false negative rates. Note that false positive and false negative rates are considered in the context of investigative objectives. In other words, false positive means retrieved data that match the query, but that are irrelevant to the investigation. Similarly, false negative means that data relevant to the investigation are not retrieved, because they did not match the query.

Investigators experience extremely high information retrieval overhead with text string searches due to the large number of hits that are irrelevant to the investigation. The average keyword search involving ten keywords or search expressions specified by an experienced investigator (one who knows which words and search expressions are more prone to high false positive rates) on the average hard drive can easily result in tens of thousands or hundreds of thousands of hits.

Investigators also experience poor investigative recall due to the use of literal search techniques. Relevant information is simply not retrieved by text string searches if the investigator's query does not cover all necessary words and/or strings. Given the tendency to limit "keyword list" length, and the effects of synonymy and vocabulary differences between individuals, there is a high likelihood that relevant evidence will not be located by text string searches in many investigations.

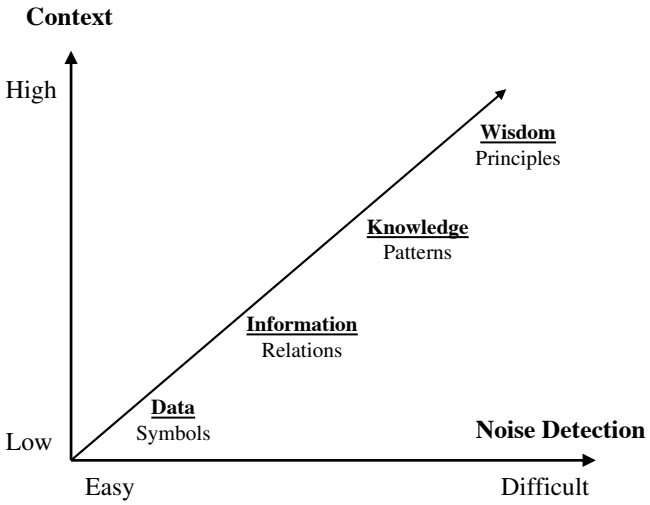

Figure 1. Knowledge Management System (KMS) Understanding Hierarchy [17].

In addition to high false positive rates and poor investigative recall, state-of-the-art digital forensic search tools do not prioritize the hits for the analyst or investigator. The hits are often ordered for presentation based on where they were found on the digital media. At most, the hits are grouped by search string and/or "file item." None of these approaches, however, appreciably reduce the information retrieval overhead. The human analyst or investigator is forced to exhaustively review all the hits to find information pertinent to the investigation, because relevant information could be listed anywhere in the output.

Some analysts and investigators use heuristics to efficiently scan the list of hits and prioritize their analytical efforts. However, such approaches are still performed manually and are relatively inefficient when compared with potential computer information processing techniques.

Indeed, if we consider digital investigations from a system theory point of view, it becomes clear that the current text string search process unnecessarily constrains the overall digital forensic investigation "system."

3. System Analysis

The purpose or mission of the digital forensic system, in the context of a single investigation, is to convert data to knowledge in accordance with the Understanding Hierarchy proposed by Nunamaker, *et al.* [17], which is presented in Figure 1.

The following terms are defined by adapting the definitions of Nunamaker, *et al.* [17] to the digital forensics context:

- **Data:** "...raw symbols that merely exist and have no significance beyond that existence...raw data lacks context and therefore does not have meaning in and of itself nor does it have any meaningful relations to anything else." In the digital forensic system, data refers to the raw data stored on digital media before meaning is extracted from it (i.e., the binary data stream).

- **Information:** "...data organized into meaningful relationships and structures...Information is factual in nature and conveys description..." In the digital forensic system, information refers to data at a higher level of abstraction [2], wherein the data is represented in a form more meaningful to human investigators and analysts, but from which knowledge has yet to be extracted (i.e., representing binary data as a human-readable word).

- **Knowledge:** "...information organized into meaningful patterns and repeatable processes...The patterns develop into knowledge when someone has an awareness and understanding of the pattern and its implications." Knowledge is synthesized information and information in context. It enables the user of information to apply it or to make use of it. In digital forensics, knowledge refers to the synthesis of information by human analysts to attain the investigative goals; it is the set of answers to investigative questions posed.

- **Wisdom:** "...tends to be self-contextualizing...[it] is knowledge organized into principles we use to reason, discern or judge." Wisdom refers to the highest level of understanding. In digital forensics, wisdom is most appropriately defined as the knowledge an investigator takes away from a specific investigation that can enable better (i.e., more effective and/or more efficient) digital forensic investigations in the future.

Given these definitions and the fact that the purpose of a digital forensic investigation is to confirm or refute allegations and/or obtain intelligence information, it becomes clear why the purpose of the digital forensic "system" is to convert data to information and information to knowledge.

The conversion from data to information in the digital forensic system is a two-step process. First, as previously described, a text string search produces a list of hits. Second, the human analyst or investigator reviews the list and creates a prioritized list of hits to guide subsequent analytical efforts. Both the original list of hits and the manually prioritized list of hits represent information. The former is created by

computer information processing (CIP) and the latter by human information processing (HIP). The conversion from information to knowledge is facilitated by HIP, as the basic definition of knowledge necessitates. Specifically, the human analyst or investigator reviews the hits, interrogates the raw data and/or information as needed, and draws conclusions related to the investigation. These conclusions represent knowledge.

The information to knowledge conversion process, which uses standard digital forensic analysis tactics, techniques, tools and procedures, is laborious and inefficient. This is because of the high information retrieval overhead and the voluminous output produced by the data to information conversion process. The inefficiency is so stark that investigators often limit the scope of their text string searches to a single search, comprising a keyword or search expression list that is shorter than desired (ten search expressions or less). This action severely constrains the data to information conversion process – some information (in the form of search hits) is not retrieved because of the limited search query. Fundamentally, this constraint runs contrary to the purpose of the digital forensic system.

To illustrate the point that the data to information conversion process is overly constrained (arguably, even artificially constrained), consider a search query issued in a murder investigation. If it is alleged that the suspect sent a web-based email to the victim, wherein she threatened to kill the deceased, the investigator or analyst might wish to search the hard drive for the word "kill." An experienced digital forensics investigator, however, would caution against the inclusion of the word "kill," because of the extremely high number of false positive hits that would result (with respect to investigative objectives). The high false positive rate occurs because the word "kill" is present in various system commands, processes and documentation. Omission of the word "kill" from the search query then becomes a constraint on the data to information conversion process, as certain web-based email evidence may be missed unless it is found by another keyword search or by some other search technique.

The data to information conversion process is also unnecessarily constrained by vocabulary differences – there is only a 20% probability that two people will use the same search terms given the same search objective [7]. For example, a literal search that includes "Saddam Hussein" but not "Iraq" will only retrieve hits containing "Saddam Hussein." It will not retrieve relevant hits that contain the word "Iraq" but do not contain the expression "Saddam Hussein." Non-literal search techniques would retrieve both sets of hits. Thus, literal search techniques overly constrain the data to information conversion process.

A closer examination of the data to information conversion process reveals that it becomes overly constrained to counteract insufficient regulation of the information to knowledge conversion process. In this instance, regulation serves to control the display of information, enhancing knowledge acquisition. A regulator that prioritizes and/or groups search hits according to the probability of each hit being relevant would greatly improve the information to knowledge conversion process. Such regulators have been used by Internet search engines for more than ten years. Surprisingly, they have not yet been employed in digital forensic text string searching.

If we accept the argument that the text string search process (or subsystem) is performing sub-optimally, thereby causing the overall digital forensic investigation system to perform sub-optimally, then the natural question is: How do we improve the system's performance?

4. Process Model

The optimized text string search process (subsystem) can be achieved by *appropriately* constraining the data to information conversion process and by increasing the regulation of the information to knowledge conversion process. Currently, the amount and variety of information produced from the text string search process is overwhelming. This problem exists even in the face of an overly constrained data to information conversion process, wherein only a subset of relevant information is obtained. A regulator is needed to reduce the information retrieval overhead, thereby improving the information to knowledge conversion process.

At the same time, however, problems caused by synonymy and vocabulary differences require that the data to information conversion process be constrained appropriately. In other words, instead of simply reducing the number of permissible search strings (which exacerbates the synonymy and vocabulary differences problems), it is important to handle search strings more intelligently and, possibly, even expand the scope (quantity) of permissible strings.

The proposed text string search process model introduces additional system state transitions and operators to address the problems posed by high information retrieval overhead and poor investigative recall. The two primary types of operators in the text string search process are computer information processing (CIP) operators and human information processing (HIP) operators. As discussed above, the current data to information to knowledge conversion process is a three-step process involving only one CIP approach. The proposed process model involves

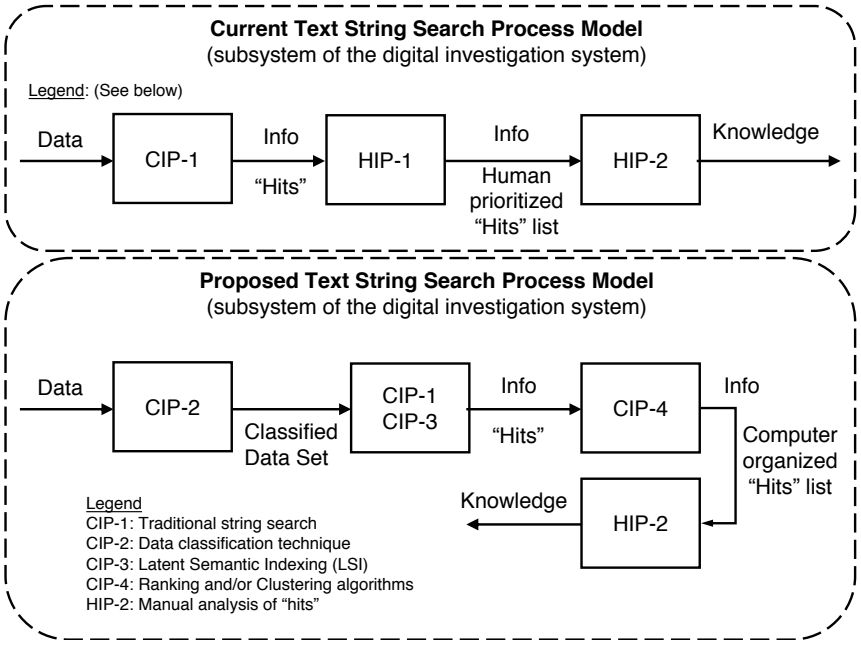

Figure 2. Current and proposed text string search process models.

a four-step process with four advanced CIP approaches and only one HIP approach (Figure 2).

The first step in the proposed text string search process model involves data classification. The goal of this step is to enable filtering so that successive CIP processes may be executed against smaller data subsets. This is important given the increased computational expense that will likely accompany the introduction of additional CIP processes. For example, data may be categorized as text or non-text by a CIP data classification technique such as Shannon's Forensic Relative Strength Scoring System (FRSS) [22], which uses two measures (ASCII proportionality and entropy) to categorize data as text or non-text.

The second step of the text string search process model involves the application of two CIP approaches in concert with each other. The goal of this step is to generate a set of search hits that is more exhaustive (for investigative purposes) than that produced by traditional search techniques. The step also reduces the effects of the synonymy and vocabulary differences problems. The first CIP process is traditional text string searching, which is augmented with a latent semantic indexing (LSI) technique to facilitate the extraction of semantic meaning from text [24]. Latent semantic indexing (LSI) was pioneered in 1990 by

Deerwester, *et al.* [6]. Several variants and improvements have since been developed. Among the most notable is probabilistic latent semantic analysis (PLSA or pLSI), which was introduced by Hofmann [11]. PLSA improves the quality of the semantic dimensions by including and excluding terms from the semantic subspace based on statistical significance rather than quantitative thresholds alone.

In this application, LSI serves to reduce false negatives by allowing hits that are relevant to the investigation to be retrieved; these hits would not otherwise be retrieved via traditional literal search techniques. Using LSI to reduce false negatives also improves the quality of information generated during the data to information conversion process. Such computer information processing techniques have been applied successfully in other areas, including web content mining [5] and semantic extraction from on-line postings [19].

The third step of the text string search process model involves another (new) CIP approach – the application of relevancy ranking and/or clustering algorithms. The goal is to reduce the effects of polysemy. A relevancy ranking algorithm serves to prioritize the hits (i.e., order the hits according to investigative relevance), thereby reducing information retrieval overhead and helping investigators get to the relevant hits faster. As stated previously, this approach is widely employed by Internet search engines, but it has not been used in the digital forensics arena to improve text string searching. This is largely due to the fact the algorithms are not directly applicable to digital forensic text string searching given the variables used by the algorithms and the nature of the data sets. Internet search engine prioritization ranking algorithms rely heavily on web-oriented variables such as PageRank, anchor text and visual properties of the text [1], which are not relevant to digital forensic data sets. Other ranking variables, e.g., proximity measures and query term order, are probably extensible, but research has shown that these variables alone are insufficient to achieve good results.

The basic premise of post-retrieval clustering is that information retrieval overhead can be significantly reduced by thematically grouping query results, thereby enabling the user to find relevant hits more efficiently. This is due to the "cluster hypothesis," which states that computationally similar documents tend to be relevant to the same query [25]. Empirical research has shown that clustered query results improve information retrieval effectiveness over traditional ranked lists using static relevance ranking models. Such results hold true for both traditional text-based information retrieval (i.e., clustering retrieved documents from a digital library) [9, 12–15] and web-based information retrieval [26–28].

The final step of the process model is unchanged from the current process model: HIP-2, the manual analysis of search results. It is postulated that HIP-1, manual prioritization heuristics, will no longer be necessary, and that HIP-2 will be more effective and more efficient. The introduction of additional computer information processing steps will improve investigative recall and decrease information retrieval overhead. We expect that the increased CIP time associated with the additional CIP steps will be greatly overshadowed by the gains in effectiveness and efficiency due to decreased HIP demands.

5. Conclusions

The proposed process model represents a paradigmatic change in digital forensic text string searching. It is admittedly theoretical and high-level in nature, but it should stimulate new ways of thinking about digital forensic text string searching. Candidate classification, LSI, ranking and clustering algorithms must be identified and empirically tested to assess query precision, query recall and computational expense. Similar algorithms have already been developed for information retrieval, artificial intelligence, data mining and text mining applications. The challenge is to redesign these algorithms to improve text string searching in the area of digital forensics.

The process model has two principal benefits. First, the effectiveness and efficiency of text string searching in digital investigations will be greatly improved. The data to information conversion process will no longer be artificially constrained, and the information to knowledge conversion process will be enhanced. In short, less evidence will be missed, and the human analysis portion of the search process will become much less laborious.

The second benefit is the potential scalability to large data sets. Traditional string matching approaches are impractical for terabyte data sets – the overall process time is inordinately large given the number of false positives and the resulting information retrieval overhead. Most database and tree-indexing algorithms are also impractical when conducting a small number of searches against a terabyte data set. Again, the problem is inordinately high information retrieval overhead. In this instance, however, the problem is not just HIP time, but also excessive CIP time, given the amount of indexing time required before an investigation can begin (and the limited return on investment as the index creation time is not spread across multiple searches).

The proposed approach has certain limitations, including (possibly) high error rates, computational expense and the difficulty of extending

data mining algorithms to digital investigations. Error rates have been largely ignored in digital forensics research [3]. An understanding of error rates becomes even more important with the proposed model because of the use of advanced CIP approaches. The data classification step inherently filters the data set leading to concerns that not everything is being searched. The use of ranking and/or clustering algorithms might dissuade investigators from exhaustively reviewing the search results. Moreover, investigators might even cease their reviews when "enough" evidence is obtained.

Clearly, additional and more computationally expensive CIP algorithms will increase computer processing time. The fundamental question, however, is whether or not the processing time is dwarfed by the HIP time that is required when false positives and false negatives are not controlled, and when results are not prioritized and/or summarized.

Finally, the extensibility of data mining algorithms to digital investigations is a major concern. On the one hand, much of the research in text mining and web content mining has been driven by the demand for eliciting "business intelligence" from massive data sets. On the other hand, raw binary data on digital media is unlike the data sources for which data mining algorithms were developed. Indeed, in digital forensic applications, data sources are likely to be very heterogeneous and unstructured.

The issues related to error rates, computational expense and data mining algorithm extensibility must be carefully considered when candidate algorithms are selected and tested. Nonetheless, it is expected that the proposed process model will drastically decrease false positive and false negative error rates, improve query recall and precision, and reduce the human analytical burden – all while allowing the investigator to increase the number of search terms used on terabyte or larger data sets.

References

[1] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems*, vol. 30(1-7), pp. 107–117, 1998.

[2] B. Carrier, Defining digital forensic examination and analysis tools using abstraction layers, *International Journal of Digital Evidence*, vol. 1(4), pp. 1–12, 2003.

[3] E. Casey, Error, uncertainty and loss in digital evidence, *International Journal of Digital Evidence*, vol. 1(2), pp. 1–45, 2002.

[4] E. Casey, Network traffic as a source of evidence: Tool strengths, weaknesses and future needs, *Digital Investigation*, vol. 1, pp. 28–43, 2004.

[5] S. Das and M. Chen, Yahoo! for Amazon: Opinion extraction from small talk on the web, *Proceedings of the Eighth Asia-Pacific Finance Association Annual Conference*, pp. 1–45, 2001.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol. 41(6), pp. 391–407, 1990.

[7] G. Furnas, L. Gomez, T. Landauer and S. Dumais, The vocabulary problem in human-system communication, *Communications of the ACM*, vol. 30, pp. 964–971, 1987.

[8] J. Giordano and C. Maciag, Cyber forensics: A military operations perspective, *International Journal of Digital Evidence*, vol. 1(2), pp. 1–13, 2002.

[9] M. Hearst and J. Pedersen, Reexamining the cluster hypothesis: Scatter/gather on retrieval results, *Proceedings of the Nineteenth ACM International Conference on Research and Development in Information Retrieval*, pp. 76–84, 1996.

[10] F. Hinshaw, Data warehouse appliances: Driving the business intelligence revolution, *DM Review Magazine*, September 2004.

[11] T. Hofmann, Probabilistic latent semantic indexing, *Proceedings of the Twenty-Second ACM International Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[12] A. Leuski, Evaluating document clustering for interactive information retrieval, *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 33–40, 2001.

[13] A. Leuski and J. Allan, Improving interactive retrieval by combining ranked lists and clustering, *Proceedings of the Sixth RIAO Conference*, pp. 665–681, 2000.

[14] A. Leuski and J. Allan, Interactive information retrieval using clustering and spatial proximity, *User Modeling and User-Adapted Interaction*, vol. 14(2-3), pp. 259–288, 2004.

[15] A. Leuski and W. Croft, An evaluation of techniques for clustering search results, Technical Report IR-76, Computer Science Department, University of Massachusetts at Amherst, Amherst, Massachusetts, pp. 1–19, 1996.

[16] S. Mukkamala and A. Sung, Identifying significant features for network forensic analysis using artificial intelligence techniques, *International Journal of Digital Evidence*, vol. 1(4), pp. 1–17, 2003.

[17] J. Nunamaker, N. Romano and R. Briggs, A framework for collaboration and knowledge management, Proceedings of the Thirty-Fourth Hawaii International Conference on System Sciences, 2001.

[18] D. Radcliff, Inside the DoD's Crime Lab, *NetworkWorldFusion*, pp. 1–5, March 8, 2004.

[19] B. Rajagopalan, P. Konana, M. Wimble and C. Lee, Classification of virtual investing-related community postings, *Proceedings of the Tenth Americas Conference on Information Systems*, pp. 1–6, 2004.

[20] V. Roussev and G. Richard, Breaking the performance wall: The case for distributed digital forensics, *Proceedings of the Fourth Annual Digital Forensics Research Workshop*, pp. 1–16, 2004.

[21] M. Schwartz, Cybercops need better tools, *Computerworld*, p. 1, July 31, 2000.

[22] M. Shannon, Forensics relative strength scoring: ASCII and entropy scoring, *International Journal of Digital Evidence*, vol. 2(4), pp. 1–19, 2004.

[23] P. Sommer, The challenges of large computer evidence cases, *Digital Investigation*, vol. 1, pp. 16–17, 2004.

[24] D. Sullivan, *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales*, Wiley, New York, p. 542, 2001.

[25] C. van Rijsbergen, *Information Retrieval*, Butterworths, London, 1979.

[26] O. Zamir and O. Etzioni, Web document clustering: A feasibility demonstration, *Proceedings of the Twenty-First ACM International Conference on Research and Development of Information Retrieval*, pp. 46–54, 1998.

[27] O. Zamir and O. Etzioni, Grouper: A dynamic clustering interface to web search results, *Computer Networks*, vol. 31(11-16), pp. 1361–1374, 1999.

[28] H. Zeng, Q. He, Z. Chen, W. Ma and J. Ma, Learning to cluster web search results, *Proceedings of the Twenty-Seventh ACM International Conference on Research and Development in Information Retrieval*, pp. 210–217, 2004.